

AD_____

Award Number: DAMD17-02-1-0406

TITLE: Molecular Differentiation of Risk for Disease Progression: Delineating Stage-Specific Therapeutic Targets for Disease Management in Breast Cancer

PRINCIPAL INVESTIGATOR: Maria J. Worsham Ph.D.
Usha Raju M.D.
Mei Lu Ph.D.

CONTRACTING ORGANIZATION: Henry Ford Health System
Detroit, Michigan 48202

REPORT DATE: July 2006

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE				<i>Form Approved</i> OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) 01-07-2006		2. REPORT TYPE Final		3. DATES COVERED (From - To) 1 Jul 2002 – 30 Jun 2006	
4. TITLE AND SUBTITLE Molecular Differentiation of Risk for Disease Progression: Delineating Stage-Specific Therapeutic Targets for Disease Management in Breast Cancer				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER DAMD17-02-1-0406	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Maria J. Worsham Ph.D.; Usha Raju M.D. and Mei Lu Ph.D. E-Mail: mworsha1@hfhs.org				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Henry Ford Health System Detroit, Michigan 48202				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT Cancer is a highly heterogeneous disease, both morphologically and genetically. A current shortcoming in cancer prognosis and treatment is a lack of methods that adequately address the complexity and diversity of the disease. Genome wide studies can provide molecular characterization or fingerprints of cancer phenotypes linked to clinical information. The aim of this research is to 1a: identify an informative set of specific genetic alterations that underlie the pathogenesis of disease progression to serve as targets for management of disease at the earliest stages and 1b: refine stage-specific disease phenotypes by integration of molecular profiles with known risk factors of breast cancer such as reproductive characteristics, medical history, and histologic parameters of breast carcinomas. We will examine 100 cases in each disease stage category of 0, 1, 2, 3, 4 and unknown to evaluate 120 breast cancer associated gene markers distributed throughout the human genome. Molecular fingerprints identified from genome wide studies should delineate patterns of genomic imbalances at the level of stage-specific gene loci, providing a novel index to estimate the extent of genomic abnormality with disease progression. This knowledge should allow the integration of stage-specific therapeutic targets as treatment intervention strategies in the management of breast cancer.					
15. SUBJECT TERMS disease progression, genome wide, gene loci alterations, stage-specific markers, breast cancer					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (include area code)
			UU	29	

Table of Contents

Introduction.....	4
Body.....	4
References	4
Key Research Accomplishments.....	5
Reportable Outcomes.....	6
Conclusion.....	6
Appendices.....	7

INTRODUCTION

Cancer is a highly heterogeneous disease, both morphologically and genetically (1). A current shortcoming in cancer prognostication and treatment is a lack of methods that adequately address the complexity and diversity of the disease. A detailed molecular characterization or fingerprint of cancer is an objective recently made possible by the development of several new high throughput analytical methods. These include techniques for the analysis of DNA, mRNA, and proteins within a cell (2-4). Building databases of detailed molecular information and linking them to clinical information are very attainable goals (5). This approach has the potential to help patients by improving grouping of tumor subtypes, which may enable clinicians to more accurately distinguish prognostic groups, and predict the most effective therapies. Prognostic marker systems based on single parameters have generally proven inadequate. Thus, multiparametric methods, which rely on many pieces of information, are ideally suited to the grouping of tumor subtypes and the identification of specific patterns of disease progression.

A major objective of current cancer research is to develop a detailed molecular fingerprint of tumor cells and tissues that is linked to clinical information. Toward this end, using the Multiplex Ligatable Probe Amplification technique (MLPA, 6), a novel assay recently developed at MRC Holland (Amsterdam) we will interrogate 120 gene loci (Table 1, Study Instruments) altered in breast cancer using a nested case cohort of 600 stage-specific breast cancers drawn from a retrospective cohort of 6000 primary breast cancers.

REFERENCES

1. Tavassoli, F. A., and Schnitt, S. J. Pathology of the Breast. New York: Elsevier, 1992.
2. Pollack J. R., Perou C. M., Alizadeh A. A., Eisen M. B., Pergamenschikov A., Williams C. F., Jeffrey S. S., Botstein D., Brown P. O. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat. Genet., 23: 41-46, 1999
3. Duggan D. J., Bittner M., Chen Y., Meltzer P., Trent J. M. Expression profiling using cDNA microarrays. Nat. Genet., 21: 10-14, 1999
4. Oh J. M., Hanash S. M., Teichroew D. Mining protein data from two-dimensional gels: tools for systematic post-planned analyses. Electrophoresis, 20: 766-774, 1999
5. Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., Lander E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (Washington DC), 286: 531-537, 1999
6. MRC Holland: Dr. J. P. Schouten, mrch@bio.vu.nl. Website: www.MRC-Holland.com

BODY

Statement of Work

Task 1. Cohort construction, Months 1-24

- a: Begin construction of the breast cancer study cohort. We have identified 6000 breast cancer cases in the HFHS system from 1981 through 2000. Drs. Worsham and Chase will select 100 stage-specific breast cancers corresponding to stage 0 (in situ), stage 1, 2, 3, 4 and unknown stage
- b: Set up database of study cohort

- c: Retrieval of H & E slides for cancer cohort
- d: The Pathologist Dr.Raju and the P.I will begin pathology review of the cancer cohort recording histopathological characteristics on the Pathology Cancer Review Form (see study instruments).
- e: Data entry of histopathology indicators
- f: selection of tumor blocks and sectioning of tissue for microdissection and DNA extraction

Task 2. Molecular Assays Months 3-34

- a: Begin the novel Multiplex Ligatable Probe Amplification(MLPA) assays
- b: As DNA becomes available set up molecular worksheets and forms for electronic data entry of molecular data (Teleform)

Task 3. Medical chart abstraction Months 3-32

- a: Begin medical chart abstraction using the Medical record Abstraction Form
- b: Data entry of forms into the study database

Task 4. Interim Analyses, Months 18-24

- a: Interim statistical analysis of data obtained from molecular, pathology, and medical record abstractions will be performed periodically
- b: Annual reports will be written

Task 5. Final Analyses and Report Writing, Months 32-36

- a: Final analyses of data form molecular, pathology and medical abstractions will be performed
- b: A final report and initial manuscripts will be prepared

PROGRESS (July 1, 2003- June 30, 2004:

KEY RESEARCH ACCOMPLISHMENTS

Task 1 accomplishments:

• **July 1, 2002-June 30th 2006:**

We have so far acquired a total breast cancer patient database of 5008 validated and verified breast cancer cases. The study cohort of 600 stage-specific breast cancer subjects was derived from this comprehensive patient database. Selection of breast cancer subjects in each of the 6 stages, stage 0, stage 1, stage 2, stage 3, stage 4, and stage unknown was performed by the biostatistician Dr. Mei Lu (Dr. Lu replaced Dr. Chase). Criteria for selection were as follows: 1) age <50 years, Caucasian American (CA); age <50 years, African American (AA); 2) age >50 years, CA; age >50 years, AA. A total of 1,244 subjects were obtained as a result of this selection; stage 0=215; stage 1= 225; stage 2= 228; stage 3= 188; stage 4= 179; stage unknown= 209. Further selection of 100 stage-specific cases for equal representation of CA and AA in each of the two age categories, <50 years and > 50 years was achieved in a random fashion by Dr. Gary Chase. Thus, the study cohort of 600 breast cancer subjects, 100 in each of the 6 stages has been completed. **Status: Completed**

July 1, 2002-June 30th 2006

- Data bases of the study cohort have been completed and linked with the Henry Ford Health System Tumor Registry for demographics, histopathology, and clinical information. The latter has been obtained for the entire cohort of 1,244 subjects. **Status: Completed**

July 1, 2002-June 30th 2006:

- The study Pathologist Dr. Raju has completed review of 560 breast cancer subjects (missing slides for review from the Pathology archives accounted for 40 cases)

Status: Completed

July 1, 2002-June 30th 2006

- Pathology Review Form data via electronic Teleform data entry.

Status: Completed

July 1, 2002-June 30th 2006

- Tissue block retrieval, sectioning, H & E staining, microdissection, and DNA extraction has been accomplished for 465 (missing or unavailable tissue blocks account for approximately 135 cases)

Status: Completed

Task 2**July 1, 2002-June 30th 2006**

Multiplex Ligatable Probe Amplification (MLPA) assays have been performed for a total of 425 Cases (insufficient DNA, missing or unavailable blocks accounted for 135 cases). 145 patients. **Status: Completed**

Task 3**July 1, 2002-June 30th 2006**

- Medical record abstraction has been completed for 390 subjects and entered into the database
- Status: Completed**

Task 4: Interim Analyses

Status: Interim statistical analysis of 263 breast cancer patients for a total of 595 records was performed and present at the June 2005 Era of Hope meeting (**Appendix item 1 & 2**)

Task 5. Final Analyses and Report Writing, Months 32-36

Final analyses of data for molecular, pathology and medical abstractions: **Status: Completed**
Manuscripts:

- 1: “Molecular classification of breast carcinoma in situ” (Appendix item 3)
- 2: “Molecular modeling tumor versus non-tumor, in preparation
- 3: Stage specific genetic algorithms in breast cancer, in preparation

CONCLUSION/REPORTABLE OUTCOMES

1: Worsham MJ, Yang J, Tiwari N, Chen KM, Cheng J, Shah V, Raju U. Delineating a molecular continuum for breast cancer progression: molecular modeling individual gene loci alterations in breast cancer”. DOD Era of Hope, Philadelphia, PA, June 8th, 2005

2: Raju U, Sethi S, Lu M, Qureshi H, Cheng J, Yang J, Tiwari N, Chen KM, Worsham MJ. Molecular differentiation of breast carcinoma in situ. Symposium presentation, DOD Era of Hope, Philadelphia, PA, June 8th, 2005

3: Raju U, Lu M, Sethi S, Qureshi H, Wolman SR, Worsham MJ. “Molecular classification of breast carcinoma in situ” In press, November 2006, Current Genomics

Appendices

1: Delineating a molecular continuum for breast cancer progression: molecular modeling individual gene loci alterations in breast cancer. *Poster presentation, DOD Era of Hope, Philadelphia, PA, June 8th, 2005* **Abstract:**

A current shortcoming in cancer prognostication and treatment is a lack of methods that adequately address the complexity and diversity of the disease. Prognostic marker systems based on single parameters have generally proven inadequate. Thus, multiparametric methods, which rely on many pieces of information, are ideally suited to the grouping of tumor subtypes and the identification of specific patterns of disease progression.

To test the hypothesis that the extent of genomic imbalance at gene loci known to be involved in breast cancer is progressive, and is evidence for a molecular continuum for breast cancer progression, genetic alterations of loss and gain of individual gene loci were identified using a novel assay that interrogated an evidence-based panel of 122 gene loci implicated in breast cancer, many of which are distributed along critical pathways utilized by breast cancer cells. The study cohort comprised 263 breast cancer patients for a total of 595 records, inclusive of 191 normal breast epithelium records, 93 benign breast lesions, 117 carcinoma in situ lesions, and 193 tumor lesions. Tumor (T) areas concurrent with benign proliferative lesions (BPL), normal breast epithelium (N), and in situ lesions (CIS) of DCIS and LCIS within a 5 micron section were marked by the study pathologist as part of the pathology review process and individually microdissected for DNA extraction. Statistical analysis was carried out using log-linear models, specifically the uniform association (UA) model in which scores of 1 to 4 were assigned for variables of lesion type (N, BPL, CIS, T). Scores for gene copy number ranged from 0 to 5, where 0 was homozygous loss, 1 was loss of 1 copy, 2 normal gene copy number, and 3, 4, and 5 gene copy number gain.

There was excellent goodness-of-fit for the UA model and further comparison of the UA model with independent models indicated a significant association between lesion type and gene copy number alterations. Twenty four candidate genes emerged based on their ability to discriminate among the four lesions types of N, BPL, CIS and T. A subset analysis, performed to test the ability the 24 candidate genes to distinguish between specific categories of N and BPL, BPL and CIS, and CIS and T identified statistically significant unique and overlapping targets. Three genes, BCL2, MYC, and PTEN appear to initiate the normal to benign transformation process. Progression from benign to carcinoma in situ was attributed to 4 additional genes, BAX, FGF3, MSH2, and SLA, while maintaining the base line acquisition of BCL2, MYC, and PTEN (from N to BPL). In the CIS to T, CCND1 and RB1 were identified in addition to previously acquired BCL2 and SLA. Our results indicate that discrete gene loci form a key repertoire of molecules along a molecular continuum in the evolution of disease progression.

2: Molecular differentiation of breast carcinoma in situ. Symposium presentation and poster session, DOD Era of Hope, Philadelphia, PA, June 8th, 2005

Abstract

Most breast carcinomas in situ (CIS) are easily categorized as ductal (DCIS) or lobular (LCIS). A pleiomorphic variant of invasive lobular carcinoma (PILC) is known to be an aggressive variant of invasive lobular carcinoma (ILC). Because of its histological similarity and associated necrosis, most PLCIS lesions have been diagnosed as DCIS. Additional methods to adequately aid in the accurate differential diagnosis of in situ breast carcinomas has clinical implications, as current management of classic LCIS versus PLCIS and DCIS is not identical.

The overall aim of our research is to produce a multiparametric, comprehensive genome-wide molecular blueprint of CIS integrated with clinical risk factors in order to refine patient diagnosis and prognosis to aid in the clinical management of patients at the earliest disease stage.

To assess the ability to molecularly differentiate among carcinoma in situ lesions of PLCIS, LCIS, and DCIS using our genome wide strategy (DAMD17-02-1-0406), DNA from in situ breast carcinoma patients classified into the three categories of PLCIS, LCIS and DCIS were interrogated for gene loss and gain at 122 gene loci. There were 57 patients, 23 (40%) PLCIS, 13(23%) LCIS and 21(37%) DCIS. An interpretation of gene loci loss or gain was measured as the number of copies in a range of 0, 1 as loss, copy number 2 as normal, 3 or more as gain. To explore the predictive ability of gene loci to differentiate among in situ tissue categories, we compared gene loci copy number differences between two tissue categories using the two-sample t-test on a log transformation because the data lacked a normal distribution, and chi-square tests on three-category (loss, normal and gain) and two-category (normal, loss/gain) data, respectively. Thirteen gene probes with p-values <0.05 in any one of the three pair wise in situ group comparisons were noted. Among those 13 gene loci, seven gene probes were significant in proportions in at least two of the three pair wise in situ group comparisons based on the log transformation data, namely PTEN, CASP1, ING1, TNF2, NFKB1A, B2M, and CDH1, where CASP1 differed significantly in means among all of the three pair wise in situ categories. Gene locus TNFRSF1B at 1p36.3 had a higher copy number for PLCIS compared DCIS with a mean (STD) of 2.32 and 1.76 respectively, and a p-value =0.010. The absolute correlation coefficient was in a range of 0 to 1. Gene probes were highly correlated if the absolute correlation coefficient is over 0.70. The correlations among those 13 genes ranged from high for RENT2 and LMO2 or CDH1 ($r=0.68$, 0.64), to low between TNFRSF1B and ING1 ($r=0.01$). Independent validation of these molecular fingerprints will permit a more robust differential diagnosis of PLCIS aiding in the refinement of this disease phenotype as distinct from other in situ lesions.

3: "Molecular classification of breast carcinoma in situ" in press, November 2006 issue of Current Genomics.

MOLECULAR CLASSIFICATION OF BREAST CARCINOMA IN SITU

Usha Raju, Mei Lu, Seema Sethi, Hina Qureshi, Sandra R. Wolman, Maria J. Worsham

Correspondence and request for reprints: Maria J. Worsham Ph.D., FACMG., Department of Otolaryngology-Head and Neck Surgery, Henry Ford Health System, 1 Ford Place, 1D, Detroit, MI 48202
Tel: (313) 874-3350, Fax: (313) 874-1079, email: mworsha1@hfhs.org.

Key Words: Carcinoma in situ (CIS); Ductal carcinoma in situ (DCIS); Pleomorphic lobular carcinoma in situ (PLCIS); Lobular carcinoma in situ (LCIS); Classification Regression Tree (CART); molecular classification; Multiplex Ligation-dependent Probe Amplification Assay (MLPA); differential diagnosis

The manuscript was presented at the American Society of Investigative Pathology, April 17-21, 2004, Washington DC, United States

ABSTRACT

Pleomorphic variant of invasive lobular carcinoma (PILC) is an aggressive variant of invasive lobular carcinoma (ILC). Its *in situ* counterpart, pleomorphic lobular carcinoma *in situ* (PLCIS) is a recently described entity. Morphologically it has the typical architectural pattern of LCIS, but the neoplastic cells resemble intermediate grade DCIS. Molecular signatures that distinguish PLCIS from DCIS and LCIS would provide additional tools to aid in the histopathologic classification of PLCIS as a lesion distinct from LCIS and DCIS. CIS lesions, obtained from a study cohort of 38 breast cancer patients, were divided into 18 DCIS, 14 PLCIS and 6 LCIS. DNA from microdissected archival tissue was interrogated for loss or gain of 112 breast-cancer-specific genes using the Multiplex Ligation-dependent Probe Amplification Assay (MLPA). Classification

Regression Tree (CART) analysis was employed to develop a gene-based molecular classification to distinguish or separate out PLCIS from DCIS and LCIS. Molecular classification via CART, based on gene copy number, agreed with histopathology in 34/38 CIS cases. Loss of *CASPI* was predictive of LCIS (n=4) with one misclassified PLCIS. Gain of *RELA* predicted only the LCIS classification (n=2 cases). *STK15* and *TNFRSF1B* were predictive only for DCIS with no misclassifications. Gain of *EHF* and *TNFRSF1B* and loss of *NCOA3* were predictive of PLCIS, but not without misclassification. Molecular reclassification by CART was accomplished in 4 CIS cases: 1 PLCIS was reclassified as LCIS, 1 LCIS reclassified as PLCIS, and 2 DCIS cases as PLCIS. This study provides additional rationale for molecular modeling strategies in the evaluation of CIS lesions. This diagnostic aid may serve to minimize misclassification between PLCIS and DCIS, and PLCIS and LCIS, aiding to increase accuracy in the differential diagnosis of CIS lesions.

INTRODUCTION

Most breast carcinomas *in situ* are easily categorized as ductal (DCIS) or lobular (LCIS) (**Figure 1**). However, some CIS lesions have indeterminate histological features (**Figure 2**)[1, 2]. A pleomorphic variant of invasive lobular carcinoma (PILC) is known to be an aggressive variant of invasive lobular carcinoma (ILC)[3]. Its *in situ* counterpart, (PLCIS), defined by Frost et al.[4] in 1996, has not been fully defined histologically and biologically (**Figure 3**). PLCIS, like PILC, is expected to be more aggressive than LCIS (**Figure 4**)[5]. Moreover, although classic LCIS is considered a risk marker for cancer when compared to DCIS, the clinical and biological significance of PLCIS is currently unknown[4].

The cellular morphology in PLCIS is similar to that of intermediate grade DCIS (**Figure 4**). In the past, because of the histological similarity and associated necrosis, most PLCIS lesions have been diagnosed as DCIS. Treatment strategies are different for different types of CIS. If a diagnosis of LCIS is made, the patient is followed by observation[2], whereas a diagnosis of DCIS usually leads to definitive treatment, depending on the

extent and grade of DCIS (mastectomy, lumpectomy and radiation therapy, or observation alone). Because of the expected aggressive behavior of PLCIS, it is believed that treatment similar to DCIS may be warranted.

Current management of classic LCIS versus PLCIS and DCIS is not uniform, and additional methods to aid in the differential diagnosis are likely to have clinical consequences. Gene expression of E-cadherin (EC) provides some degree of lesion sub-typing (**Figure 5**)[6-8]. Although a negative EC stain can confirm a diagnosis of classic ILC or PLCIS it cannot distinguish LCIS and ILC from PLCIS. Furthermore, positive EC staining of DCIS-ID [ID not defined or indeterminate] lesions with reduced EC staining (EC-1+) (**Figure 6**) can increase the propensity for misdiagnosis. Thus, a negative EC stain cannot unequivocally distinguish DCIS-ID from PLCIS.

Molecular fingerprinting of CIS, by integrating lesion-specific genetic targets into the differential diagnosis, has the potential to provide more accurate distinction of prognostic groups and improved therapeutic strategies. The goal of this study was to test whether a molecular classification approach using gene copy number and Classification Regression Tree (CART) models can differentiate among three types of CIS: PLCIS, DCIS and LCIS.

MATERIALS AND METHODS

Subjects

The patient cohort comprised 38 breast cancer cases with CIS lesions, either concurrent with tumor (17 cases), as single CIS lesions of DCIS (9 cases), PLCIS (9 cases), and LCIS (2 cases), or in one case as concurrent CIS lesions of LCIS and PLICS, with LCIS as the lesion of inclusion (**Table 1**). All the DCIS lesions were of intermediate grade. The final CIS designation of the 38 patient cohort was as follows: 18 DCIS, 14 PLCIS, and 6 LCIS (**Table 1**).

DNA Extraction

CIS tissue and normal breast epithelium when available from each case were microdissected for DNA extraction. As a first step, 300ul of P-buffer (50mM tris-HCL, pH 8.5; 100mM NaCl, 1mM EDTA, 0.5% Triton

X100; 20mM DTT) was added to tubes containing whole 5 micron tissue sections or microdissected tissue. The tubes were heated for 15-20 min, at 90°C in a water bath and allowed to cool down to 60°C followed by the addition and mixing of 6ul of 20mg/ml Proteinase K, overlaid with 3 drops of mineral oil and spun 5 seconds at 13,000g. This was followed by a 4-16 hour (overnight) incubation at 60°C. The tubes were heated for 10min at 90°C in order to denature the Proteinase K and to disrupt nucleic acid formaldehyde adducts. Upon removal of the oil, the tubes were centrifuged for 15 min (at 13,000g) at room temperature and 250 ul of the supernatant was transferred to a clean 1.5 ml tube. After addition of 10ul 5M NaCl and 1000ml ethanol to the 250 ul supernatant, the tubes were incubated at -20°C for least 60 mins. This was followed by centrifugation for 15 mins at 13,000g, at -4°C. Upon removal of the supernatant, an additional centrifugation step for 10 secs ensured removal of the last traces of the supernatant. Finally, the pellet was air-dried and dissolved in 100ul of ddH₂O.

The Multiplex Ligation-dependent Probe Amplification Assay (MLPA)

The MLPA assay is a recent method for relative quantification of approximately 30-40 different DNA sequences in a single reaction, requiring only 20 ng of human DNA. The method has been detailed elsewhere [9-13]. The assay has been used successfully for the detection of deletions and duplications and the characterization of chromosomal aberrations for gains and losses of genes in cell lines and tumor samples [9-13]. Probes added to the samples are amplified and quantified instead of target nucleic acids. Amplification of probes by PCR depends on the presence of probe target sequences in the sample. Each probe consists of two oligonucleotides, one synthetic and one M13-derived, each hybridizing to adjacent sites of the target sequence. Such hybridized probe oligonucleotides are ligated, permitting subsequent amplification (All ligated probes have identical end sequences, permitting simultaneous PCR amplification using only one primer pair. Each probe gives rise to an amplification product of unique size between 130 and 480 bp. Probe target sequences are small (50-70 nucleotides). The prerequisite of a ligation reaction provides the opportunity to discriminate single nucleotide differences. The amplified fragments are separated on a DNA sequencer (**Figure 7**).

We have created and validated a panel of 122 breast-cancer-associated gene probes[12], distributed in 3 batches with 40, 41, and 41 probes, respectively. Normal tissue from each cancer subject serves as an internal reference when available. When normal tissue is not available from a subject, controls are obtained from breast reduction surgeries that have been reviewed and determined by the pathologist to have only normal breast epithelium. For cell lines, where normal DNA is not available, control (normal) female DNA samples are run with each probe set. Quantification of loss or gain of gene loci is determined through a process of normalization [9-13]. The latter addresses variations in the surface area of a peak (intensity) encountered due to fluctuations in the assay run, such as amount of DNA, ploidy variations, and PCR conditions. To determine gene copy number, the peak area for each probe is expressed as a percent of the total surface area of all peaks of a sample in an assay run (**Figure 7**). Relative copy number for each probe is obtained as a ratio of the normalized value for each locus (peak) of the sample to that of the normal control. A difference is significant only if the ratio is less than 0.7 (loss) or higher than 1.3 (gain). Complete loss or 0 copies is indicated by absence of a peak for that particular locus. A relative copy number of 2 is considered normal, 1 or 0 copies is considered loss, and 3 copies or more is considered gain.

Statistical Analysis

The study utilized the Classification and Regression Tree (CART[®]) analysis [14] to develop a gene-based model to discriminate among lesions in the three categories of DCIS, PLCIS, and LCIS. CART methodology, known as binary recursive partitioning, was developed in 1984 by Breiman et al., and uses non-parametric approaches [14]. The term “binary” implies that each group of patients, represented by a “node” in a decision tree, can only be split into two groups. Thus, each parent node can be split into two child nodes (**Figure 8A**). The term “recursive” refers to the fact that the binary partitioning process can be applied over and over again. Thus, each parent node can give rise to two child nodes and, in turn, each of these child nodes may

themselves be split, forming additional children (**Figure 8**). The term “partitioning” refers to the fact that the dataset is split into sections or partitioned.

CART has several advantages as a tool for data mining and predictive modeling. The tree produced represents a model or decision tree in which each node (branch) is determined by splitting the dataset on the basis of the one variable that results in the best separation as defined by values of the dependent variable (in this case, gene variables). At every branch, every variable is tested for its usefulness in further splitting. This exhaustive search for splitters can make CART computationally intensive. The relative importance of each variable is assessed based on its importance over all possible nodes and splits. In any one node, only one variable will be the best splitter although another may be a close second best (a good surrogate). The second-best variable may be a good surrogate for numerous splits without ever being selected as the best primary splitter. Its usefulness as a surrogate for multiple splits leads to its higher importance.

CART’s recursive partitioning algorithm, identifies the first gene variable with the greatest predictive power to create a first level branch (node) was applied to separating patients into three groups of PLCIS, LCIS, DCIS. It proceeded next to identify the second gene for each subgroup with the second greatest predictive power to partition patients further into the same three groups. The process was continued until no further gene was identified to achieve further classification. Finally, CART calculates the error in each category as well as the overall error. The error rate is a percentage of cases that are misclassified (e.g., a PLCIS case that is reclassified as LCIS).

To reduce the number of variables selected, we first identified a set of gene variables based on their importance from high (e.g., 100%) to low (0%) to predict lesion classification. This is followed by CART analysis on a subset of variables with a relative importance greater than 20%. The unbalanced cost ratio was used in CART to reduce the error on misclassifying a PLCIS case into the category LCIS or DCIS, or an error in misclassifying an LCIS case to the PLCIS category. To minimize error in misclassifying PLCIS as LCIS or

DCIS, and LCIS as PLCIS, we used a cost ratio of 3:1 for misclassifying PLCIS as either LCIS or DCIS, and 2:1 for misclassifying LCIS as PLCIS. For example, a 3:1 cost ratio in the partitioning of PLCIS and LCIS lesions indicates that the cost of misclassifying a PLCIS case into the LCIS category group is 3 times more egregious than the misclassification of a DCIS lesion into the PLCIS category group.

For modeling purposes, we calculated the error rate in each CIS category, as well as the error rate for the model, focusing on error reductions with respect to misclassification of a PLCIS case into either the LCIS or DCIS category, and misclassification of a DCIS case into the PLCIS category.

To avoid over-fitting the data, the leave-one-out cross-validation [14] was performed to evaluate the predictive ability when the model was applied to new data in the same patient cohort. Cross validation is a computationally-intensive method for validating a procedure for model building, which avoids the requirement for a new or independent validation dataset. In cross validation, the learning dataset is randomly split into N sections, stratified by the outcome variable of interest. This assures that a similar distribution of outcomes is present in each of the N subsets of data. One of these subsets of data is reserved for use as an independent test dataset, whereas the other N-1 subsets are combined for use as the learning dataset in the model-building procedure. The entire model-building procedure is repeated N times, with a different subset of the data reserved for use as the test dataset each time. Thus, N different models are produced, each one of which can be tested against an independent subset of the data. The remarkable fact on which cross validation is based is that the average performance of these N models is an excellent estimate of the performance of the original model (produced using the entire learning dataset) on a future independent set of patients [14].

Results:

Ten genes in the range of 29% to 100% in variable importance were selected in a univariate analysis as predictor variables from among the 122 gene probe panel (**Table 2**). The optimal tree sequence with the least error rate yielded 7 terminal nodes rate (**Table 3, Figure 8A**). The regression tree for CIS is presented in **Figure**

9. The splitting criterion for each node is given within the blue boxes. Terminal nodes (N-, red boxes) indicate class prediction based on gene copy number. The tree generated is initiated as a root node (Node 1) containing all 38 CIS cases. This node is split based on the value of a gene's copy number obtained from the list of genes determined on a univariate analysis (**Table 2**).

The parental node (Node 1) was split based on loss of *CASP1* copy number (≤ 1.5) generating terminal Node-1 and predicts a CIS class of LCIS (4 cases). This resulted in classifying 3 LCIS cases and 1 PLCIS into the LCIS class. All other CIS cases (34) become placed in Node 2. Node 2 becomes split initially through assignment of the *RELA* gene, where gain of *RELA* (gene copy > 3.5) generates terminal Node 7 and predicts only LCIS (n=2). The remaining 32 CIS cases without gain of the *RELA* gene (> 3.5) are split into Node 3, which is further split by gain of *EHF* (gene copy > 2.5) into terminal Node-6 predicting 4 PLCIS and reclassifying a LCIS as PLCIS. The remaining 27 CIS cases without gain of *EHF* become assembled into node 4 through assignment of the *NCOA3* gene loss (copy number ≤ 1.5) classifying 5 CIS as PLCIS with a resultant reclassification of 1 DCIS as a PLCIS. Node 5 CIS cases (n=22) become further split in terminal node-3 based on *STK15* copy number (< 2.5) classifying 12 CIS as DCIS without any misclassifications. The remaining 10 CIS in node 6 finally become split into terminal node 4 as a result of *TNFRSF1B* abnormal gene copy number containing only DCIS cases (n=4) and into terminal Node 5 (gene copy number < 1.5) with no misclassifications and terminal Node 6 (gene copy number > 1.5) for a PLCIS classification to include 5 PLCIS and 1 DCIS.

Four cases of CIS were misclassified; 1 PLCIS reclassified into the LCIS category, 1 LCIS reclassified as a PLCIS, and 2 DCIS cases into the PLCIS class. Error rates for LCIS, PLCIS, and DCIS were 11%, 7%, and 17%, respectively, for the learned data (**Table 4**), and 33%, 28% and 50% (**Table 5**), respectively, based on testing data (results of model validation).

DISCUSSION

Historically, the molecular pathogenesis of cancer has been examined one gene at a time. A detailed molecular characterization or fingerprint of cancer is an objective recently made possible by the development of several new high-throughput analytical methods. These include techniques for the analysis of DNA, mRNA, and proteins within a cell [15-17]. The databases of detailed molecular information can then be linked to clinical information [18]. This approach can help patients by improving classification of tumor types, enabling clinicians to distinguish prognostic groups more accurately and therefore to select the most effective therapies.

Classification and Regression Tree (CART) analysis is a statistical method to partition data sets into logically similar groups based on either numeric or categorical variables. CART produces decision trees, based on simple yes/no questions, to reveal relationships that are sometimes hidden in extremely complex datasets. CART permitted us to quantify the unique relationship between the categories of PLCIS, DCIS, and LCIS and gene copy number variables.

Several things should be pointed out regarding this CART tree. First, it is much simpler to interpret than the multivariate logistic regression model, making it more likely to be practical in a clinical setting. Secondly, the inherent “logic” in the tree is easily apparent, and it makes clinical sense. Interestingly, it has been shown that clinical decision-making rules which make sense to clinicians are more likely to be followed in clinical practice than rules in which the reasoning is not apparent.

All LCIS cases but one were correctly classified into their specific LCIS category. The misclassified LCIS was netted in terminal Node 6 as a PLCIS. A single PLCIS case (1/14) was reclassified into the LCIS category at terminal Node 1. Terminal nodes 3 and 4 correctly classified only DCIS lesions (n=12 and n=4, respectively). Two DCIS became reclassified as PLCIS through assignment of NCOA3 and TNFRSF1B gene assignments.

This study demonstrates the ability of CART analysis to predict CIS tissue types molecularly, based on gene copy number variables. Currently, PLCIS is treated like LCIS. However, the aggressive behavior and

histological pleomorphism seen in PLCIS indicate a possible association between PLCIS and DCIS that may warrant an altered clinical management. Because negative E-cadherin immunostaining does not discriminate PLCIS from LCIS, nor does it unequivocally differentiate DCIS-ID from PLCIS, additional tools would aid in the categorical classification of CIS lesions as LCIS, DCIS, or PLCIS.

The present study demonstrated a propensity for misclassification of DCIS into the PLCIS category. Their genotypic and morphological similarities add weight to consideration of PLCIS as an aggressive lesion. The study provides rationale for the utility of molecular differentiation algorithms in the evaluation of PLCIS and indeterminate CIS lesions.

The purpose of a decision tree is usually to allow the accurate prediction of outcome for future cases, based on the value of gene copy number variables. This is accomplished when a generated decision tree is saved for future use for interrogation with a new dataset to predict outcome. Because of the small sample size, and a less-than-robust validation result, a decision tree like the one generated in this study requires additional verification using an independent dataset, where cases from the new dataset are run through the tree.

From a practical standpoint, once a validated decision tree is generated, the process of CIS classification can be streamlined. Instead of starting from a 112 MLPA gene panel, a refined and focused MLPA panel comprising the 10 validated genes from the panel can provide the fluidity and practicality of an evidence-based targeted gene panel.

Acknowledgements

This study was supported by NIH CA 70923, DAMD DAMD17-00-1-0288, and DAMD17-02-1-0406 (Dr. Worsham)

References

1. Fisher ER, Costantino J, Fisher B, et al. Pathologic findings from the National Surgical Adjuvant Breast Project (NSABP) Protocol B-17. Five-year observations concerning lobular carcinoma in situ. *Cancer* 1996;78(7):1403-16.
2. Schnitt SJ, Morrow M. Lobular carcinoma in situ: current concepts and controversies. *Semin Diagn Pathol* 1999;16(3):209-23.
3. Bentz JS, Yassa N, Clayton F. Pleomorphic lobular carcinoma of the breast: clinicopathologic features of 12 cases. *Mod Pathol* 1998;11(9):814-22.
4. Frost A, Tsangaris T, Silverberg S. Pleomorphic lobular carcinoma in situ. *Case Rev* 1961:27–31.
5. Reis-Filho JS, Simpson PT, Jones C, et al. Pleomorphic lobular carcinoma of the breast: role of comprehensive molecular pathology in characterization of an entity. *J Pathol* 2005;207(1):1-13.
6. Jacobs TW, Pliss N, Kouria G, Schnitt SJ. Carcinomas in situ of the breast with indeterminate features: role of E-cadherin staining in categorization. *Am J Surg Pathol* 2001;25(2):229-36.

7. Middleton LP, Palacios DM, Bryant BR, et al. Pleomorphic lobular carcinoma: morphology, immunohistochemistry, and molecular analysis. *Am J Surg Pathol* 2000;24(12):1650-6.
8. Palacios J, Sarrio D, Garcia-Macias MC, et al. Frequent E-cadherin gene inactivation by loss of heterozygosity in pleomorphic lobular carcinoma of the breast. *Mod Pathol* 2003;16(7):674-8.
9. Worsham MJ, Pals G, Schouten JP, et al. Delineating genetic pathways of disease progression in head and neck squamous cell carcinoma. *Arch Otolaryngol Head Neck Surg* 2003;129(7):702-8.
10. Kunjoo JP, Raitanen M, Grenman S, Tiwari N, Worsham MJ. Identification of individual genes altered in squamous cell carcinoma of the vulva. *Genes Chromosomes Cancer* 2005;44(2):185-93.
11. Worsham MJ, Chen KM, Tiwari N, et al. Fine-mapping loss of gene architecture at the CDKN2B (p15INK4b), CDKN2A (p14ARF, p16INK4a), and MTAP genes in head and neck squamous cell carcinoma. *Arch Otolaryngol Head Neck Surg* 2006;132(4):409-15.
12. Worsham M, Pals G, Schouten J, et al. High-resolution mapping of molecular events associated with immortalization, transformation, and progression to breast cancer in the MCF10 model. *Breast Cancer Res Treat* 2006;96:177-86.
13. Worsham MJ, Chen KM, Meduri V, et al. Epigenetic events of disease progression in head and neck squamous cell carcinoma. *Arch Otolaryngol Head Neck Surg* 2006;132(6):668-77.
14. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. New York: Chapman & Hall (Wadsworth, Inc.), 1984.
15. Pollack JR, Perou CM, Alizadeh AA, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 1999;23(1):41-6.
16. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. *Nat Genet* 1999;21(1 Suppl):10-4.

17. Oh JM, Hanash SM, Teichroew D. Mining protein data from two-dimensional gels: tools for systematic post-planned analyses. *Electrophoresis* 1999;20(4-5):766-74.
18. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286(5439):531-7.

Table 1: Carcinoma in situ (CIS) Classification

CIS ID	Histopathology Classification	Molecular Classification (CART)	CIS ID	Histopathology Classification	Molecular Classification (CART)
Q1	DCIS	DCIS	Q22	PLCIS	PLCIS
Q2	DCIS	DCIS	Q24	PLCIS	PLCIS
Q4	DCIS	DCIS	Q25	PLCIS	PLCIS
Q5	DCIS	DCIS	Q26	PLCIS	PLCIS
Q6	DCIS	DCIS	Q27	PLCIS	PLCIS
Q7	DCIS	DCIS	Q28	PLCIS	PLCIS
Q8	DCIS	DCIS	Q29	PLCIS	PLCIS
Q9	DCIS	DCIS	Q30	PLCIS	PLCIS
Q10	DCIS	DCIS	Q51	PLCIS	PLCIS
Q13	DCIS	DCIS	Q56	PLCIS	PLCIS
Q14	DCIS	DCIS	Q57	PLCIS	PLCIS
Q16	DCIS	DCIS	Q59	PLCIS	PLCIS
Q46	DCIS	DCIS	Q41	PLCIS	*LCIS
Q49	DCIS	DCIS	Q35	LCIS	LCIS
Q52	DCIS	DCIS	Q40	LCIS	LCIS
Q54	DCIS	DCIS	Q42	LCIS	LCIS
Q15	DCIS	*PLCIS	Q43	LCIS	LCIS
Q53	DCIS	*PLCIS	Q50	LCIS	LCIS
Q17	PLCIS	PLCIS	Q37	LCIS	*PLCIS

*Misclassified samples by Gene Probes

Table 2: Univariate analysis for variable Importance of genes

Variable	Score	
NCOA3_D01N	100.00	
EHF_D01N	86.41	
TNFRSF1B_D01	75.15	
RELA_D01N	73.93	
DCC_D02N	70.68	
CASP1_D01N	66.88	
KLK3_D02N	58.74	
STK15_D01N	51.74	
RENT2_D01N	44.98	
FGF3_D01N	28.98	

Table 3: Tree Sequence

Tree Number	Terminal Nodes	Cross-Validated Relative Cost	Resubstitution Relative Cost	Complexity
1**	7	0.560 ± 0.131	0.175	-1.000
2	5	0.754 ± 0.126	0.325	0.050
3	4	0.944 ± 0.108	0.440	0.077
4	3	1.095 ± 0.073	0.583	0.095
5	2	1.067 ± 0.071	0.750	0.111
6	1	$1.000 \pm 9.16E-005$	1.000	0.167

* Minimum Cost

** Optimal

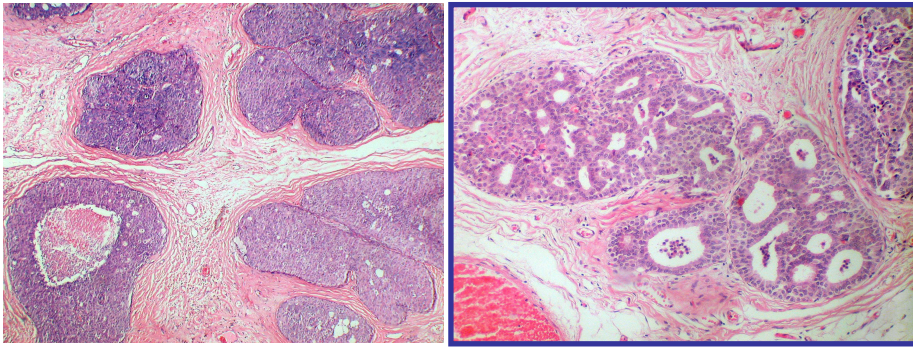
Table 4: Misclassification for Learned Data

Class	Number of Cases	Number Misclassified	Percent Error	Cost
2	18	2	11.11	0.11
3	14	1	7.14	0.07
4	6	1	16.67	0.17

Table 5: Misclassification for Test Data

Class	Number of Cases	Number Misclassified	Percent Error	Cost
2	18	6	33.33	0.33
3	14	4	28.57	0.29
4	6	3	50.00	0.50

DCIS



LCIS

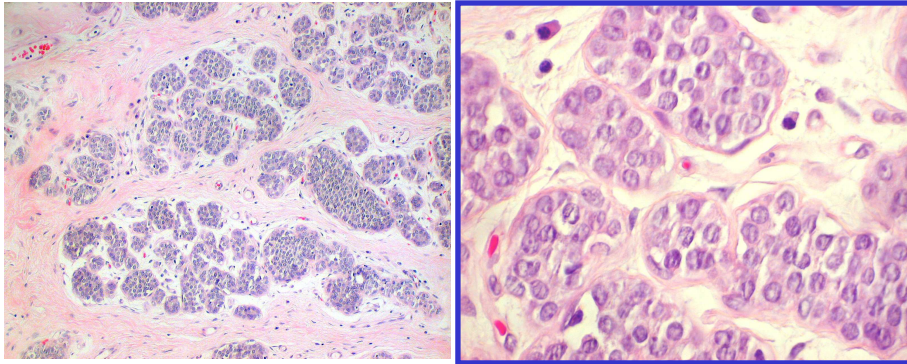
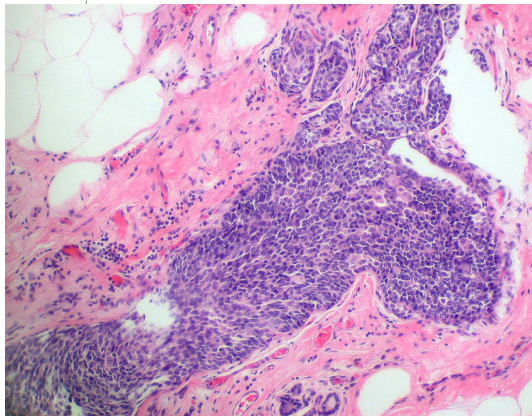
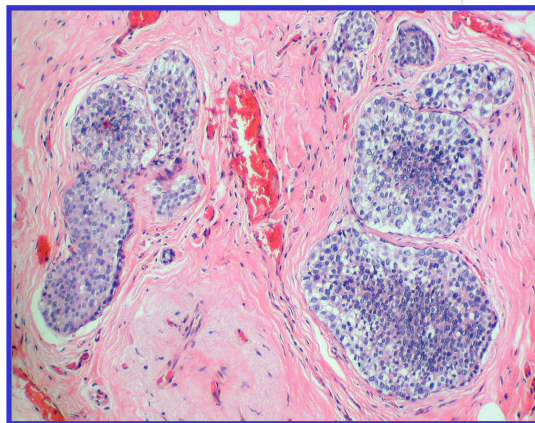


Figure 1: Most breast carcinoma in situ lesions are easily categorized as ductal (DCIS) or lobular (LCIS)

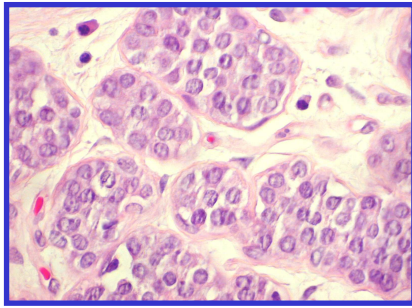


DCIS ID

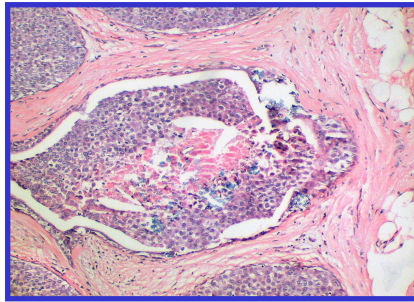


DCIS ID

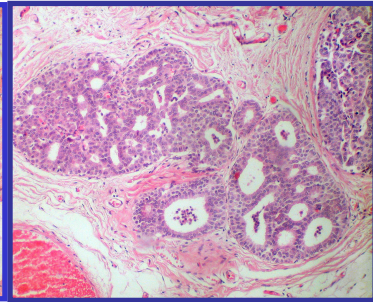
Figure 2: Some carcinoma in situ lesions have indeterminate (ID) histological features



LCIS

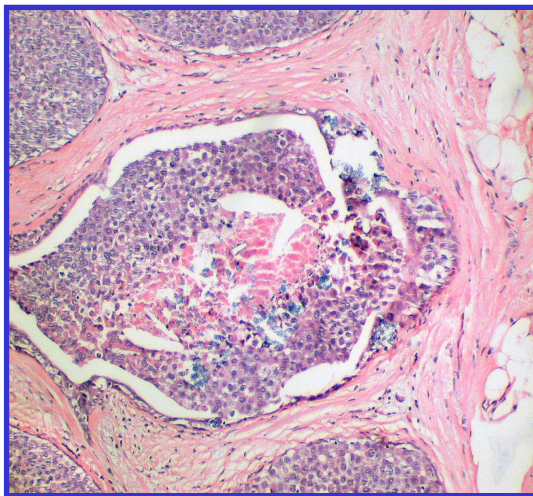


PLCIS

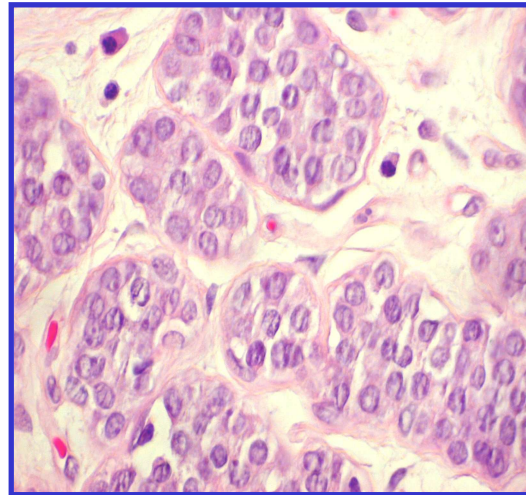


DCIS IG

Figure 3: Morphologically PLCIS has a typical architectural pattern of LCIS but the neoplastic cells resemble intermediate grade DCIS (DCIS IG).



PLCIS



LCIS

Figure 4: PLCIS is expected to be more aggressive than LCIS.

Figure 5: E-cadherin (EC) expression provides some degree of lesion sub typing. A: H &E; B: EC stain

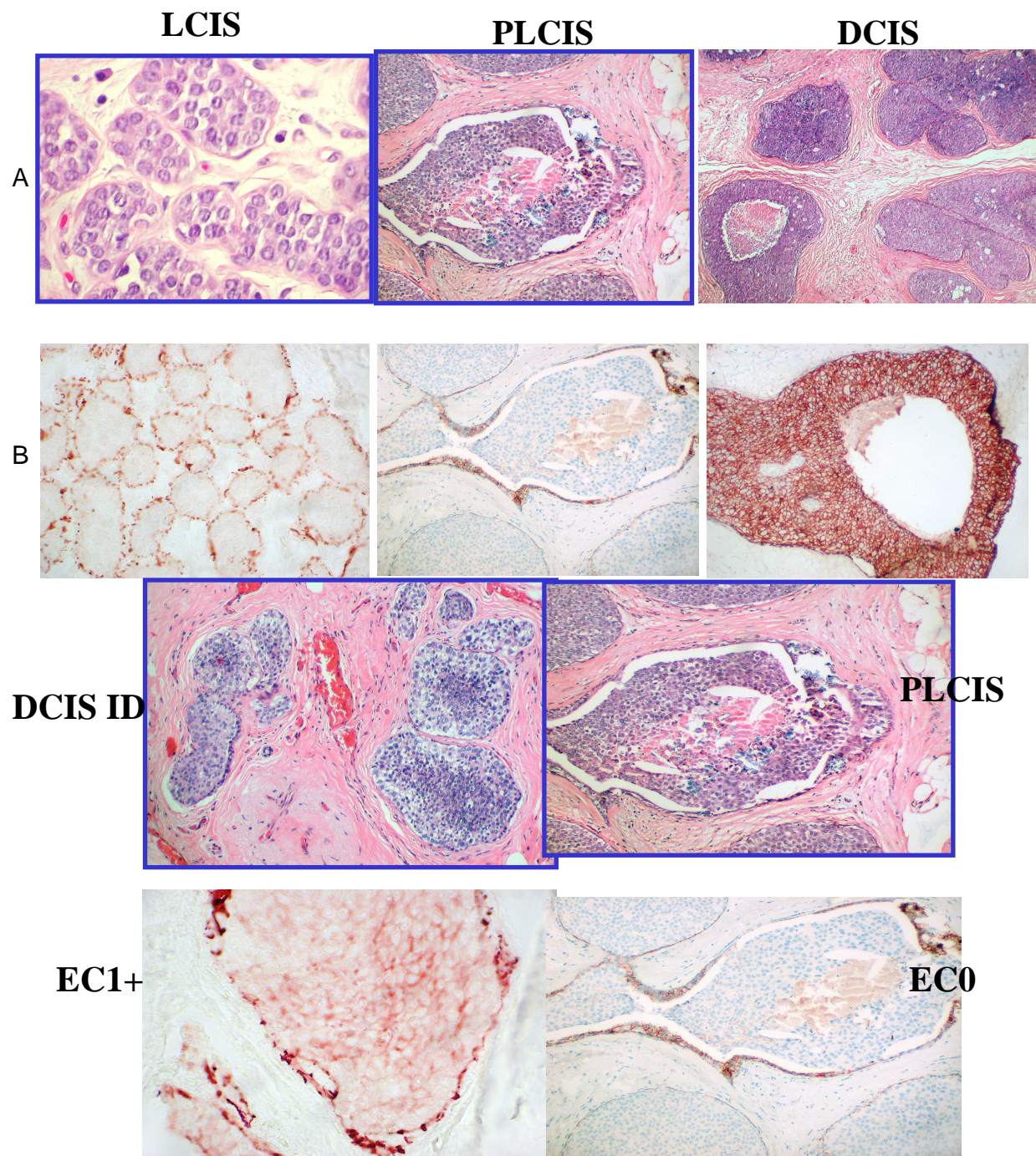
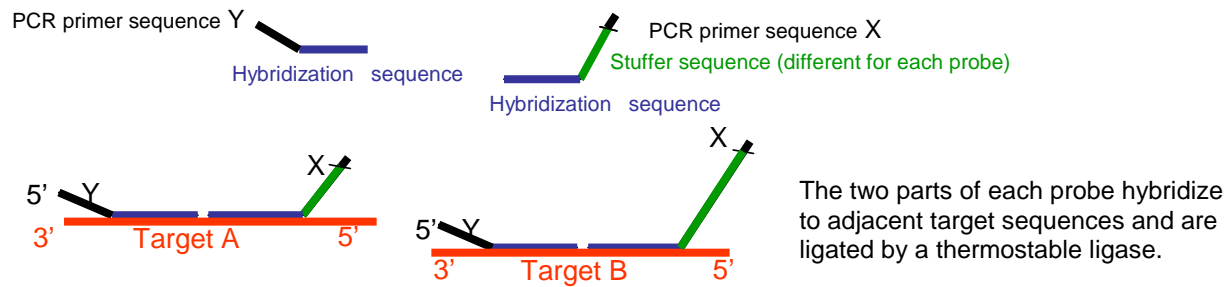


Figure 6: Limitations to EC staining: A negative EC stain cannot unequivocally distinguish DCIS ID from PLCIS.

Figure 7: Multiplex Ligation-dependent Probe Amplification (MLPA)

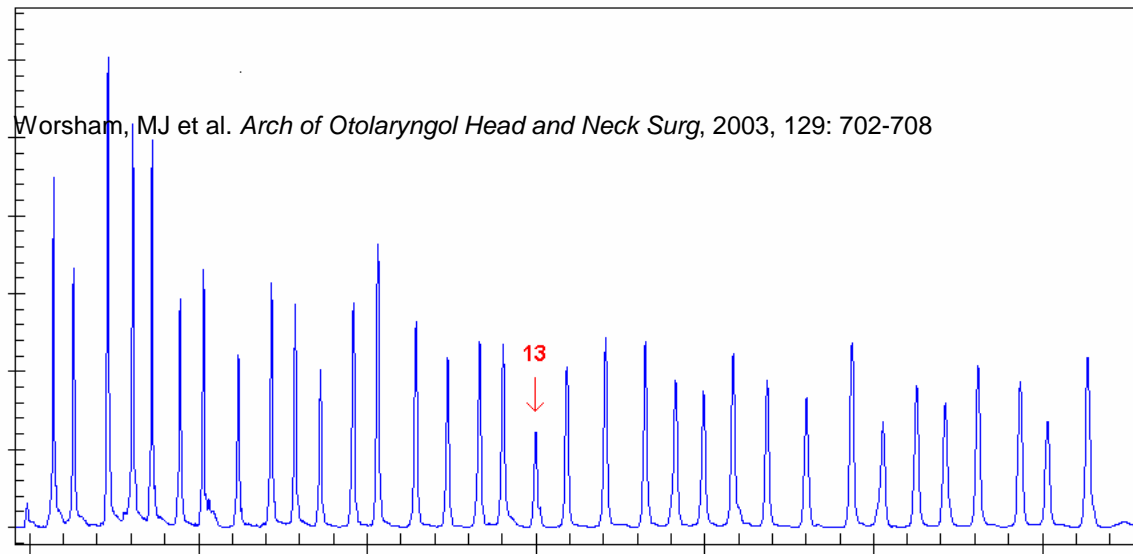
- Denatured genomic DNA is hybridized with a mixture of ~ 40 probes.
- Each MLPA probe consists of two oligonucleotides, one synthetic and one M13-derived.



All probe ligation products are amplified by PCR using only one primer pair.



Amplification products are separated on a DNA sequencer.



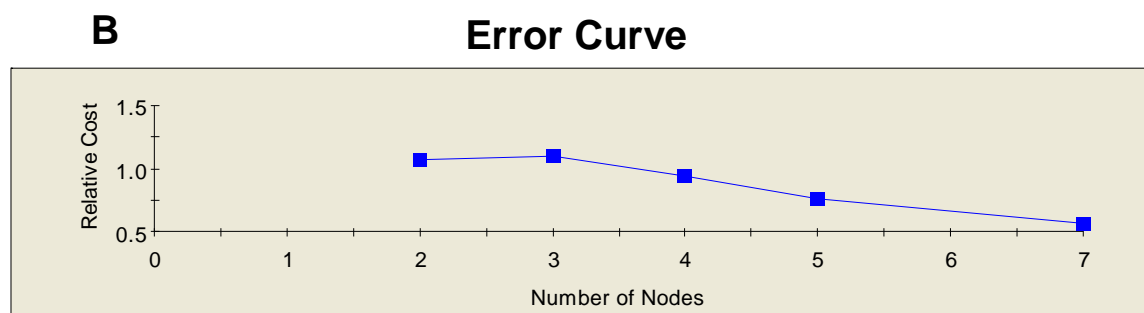
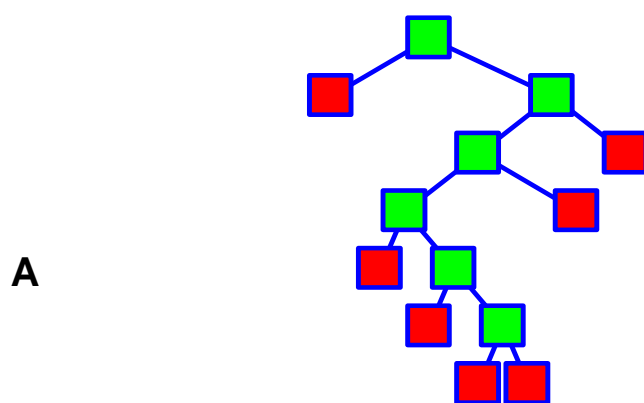


Figure 8: The optimal tree sequence with the least error rate yielded 7 terminal nodes (A) with the smallest error rate (B).

Figure 9: Regression tree for CIS. The splitting criterion for each node is given within the blue boxes. Terminal nodes (red boxes) indicate class prediction based on gene copy number. The tree generated is initiated as a root node (Node 1) containing all 38 CIS cases. This node is split based on the value of a gene's copy number obtained from the list of genes determined on a univariate analysis (Table 2).

